PREMIER MINISTRE
Liberté
Égalité
Fraternité

Direction interministérielle du numérique

Marie-Anne BARBAT- LAYANI

Paris, Le 20/11/2020

Secrétaire générale des ministères économiques et financiers

Copie
Bruno LATOMBE
Délégué aux systèmes d'information du Secrétariat
Général des ministères économiques et financiers

Yannick GIRAULT
Directeur du SCN Cap Numérique

Objet : **Avis conforme favorable - Projet « Datalake et APIM » : Valorisation et mise à disposition des données»**

Ref :

- Décret n° 2019-1088 du 25 octobre 2019 relatif au système d'information et de communication de l'Etat et à la direction interministérielle du numérique
- Courrier de saisine du 20 octobre référence DSI/2020/10/5513 reçu le 22 octobre 2020

En application de l'article 3 du décret cité en référence, vous m'avez saisi le 20 octobre 2020 pour avis sur le projet de valorisation et mise à disposition des données « Datalake et APIM ».

Ce projet vise à mettre en œuvre, pour la DGFIP, la loi pour une République numérique du 16 octobre 2016 qui a créé l'obligation pour les organisations publiques de publier et d'échanger leurs bases de données, sous réserve notamment d'anonymisation quand il s'agit de données personnelles, de protection de la propriété intellectuelle, ou du secret industriel et commercial. Ces données doivent ainsi pouvoir être exploitées et réutilisées facilement notamment par les particuliers, les entreprises et les acteurs du secteur public.

Réf : 2020-CMR-051
Affaire suivie par : Magued Abdel Maaboud / Philippe Gagnieu
Tél : 01 71 21 11 21 / 01 71 21 15 25
Mél : magued.abdel-maaboud@modernisation.gouv.fr, philippe.gagnieu@modernisation.gouv.fr

Le projet présenté « Datalake et APIM» consiste à mettre en œuvre cette démarche en s'appuyant sur des technologies complémentaires :

- Un projet « Data Lake » (ou Lac de données) qui offrira un stockage mutualisé des données brutes, qualifiées et/ou enrichies de la DGFiP permettant la simplification des mécanismes de croisements et de présentation (data visualisation) afin de pouvoir répondre plus facilement et rapidement aux demandes futures des utilisateurs. La mise en œuvre d'un dictionnaire de données et d'un module de data management fait partie de ce projet,
- Un projet « APIM » (ou API Management) qui permettra la gestion industrialisée et sécurisée des interfaces d'exposition des données de la DGFiP à l'ensemble de ses partenaires ayant juridiquement la possibilité d'y accéder en cohérence avec les exigences du RGPD.

J'ai noté que l'API Management » s'appuiera sur 3 briques logicielles interministérielles : le portail des API de l'État (api.gouv.fr) afin de promouvoir l'ensemble des API de la DGFIP, la brique d'authentification « SSO » et la brique de contractualisation « DataPass ».

Les principaux gains du projet sont :

- l'accélération et l'industrialisation des processus de partage des données DGFIP avec les partenaires externes,
- pour les agents et les services métiers de la DGFIP : le partage facilité des documents fiscaux de référence, la réalisation du profil complet des redevables particuliers et professionnels à des fins de recouvrement, la valorisation des données RH, des traitements statistiques permettant de mieux gérer les impacts de mesures gouvernementales ; etc….
- de faciliter le déploiement du programme « Dites Le Nous Une Fois »,
- sur le plan technique, ce projet permet d'améliorer la réactivité dans l'accès aux données, notamment des données des systèmes « legacy ». Il permettra également via le management des API de simplifier l'urbanisation du SI en standardisant les échanges entre applications tout en limitant les dépendances inter-applicatives,
- un gain théorique annuel de 276 ETP dans les services des impôts de particuliers (SIP) et 51 ETP hors SIP est avancé jusqu'en 2030.

Pour atteindre ces objectifs, le projet « Data Lake et APIM » mobilise 4 leviers principaux :

1. Une centralisation des données au sein d'un « Data Lake » et leur exposition au sein « d'API » permettant le décloisonnement effectif des données de la DGFiP,

2. La mise en place d'une plateforme informatique « Data Lake & API » garantissant la puissance de calcul, les logiciels et le stockage,

3. L'acquisition et la fidélisation au sein de la DGFiP de compétences de Data Science maîtrisant les dernières techniques d'analyse sur ce marché émergent,

4. La valorisation des données au travers de l'utilisation de nouvelles technologies pour diminuer les sollicitations des agents, notamment en SIP.

En l'état du dossier transmis, l'analyse du projet «Datalake et APIM» a fait remonter 6 points d'attention :

1. La fin programmée de la maintenance d'Hortonworks nécessite de revoir la trajectoire technique définie jusqu'alors par le projet. La DGFIP privilégie le développement interne d'une alternative à Cloudera. Elle est estimé à 2 M€ en développement. Les impacts calendaires et techniques de ce choix ne sont pas pris en compte sur la trajectoire affichée du projet ;
2. La trajectoire du projet Datalake apparait insuffisamment guidée par les usages métiers ce qui met en risque sa capacité à répondre à des besoins opérationnels et son appropriation à moyen terme. Le ralliement d'applications ou projets majeurs (Helios, Roc SP, Foncier Innovant….) sont évoqués, la stratégie globale de ralliement du parc applicatif actuel et cible de la DGFIP au Datalake n'est pas lisible ;
3. L'équipe cible en charge du Datalake est en cours de recrutement (sa capacité doit fortement augmenter en 2021 avec l'apport de nouvelles compétences). Les compétences au sein du projet Datalake sont donc aujourd'hui majoritairement portées par des prestataires externes ce qui pose la question de la pérennité du savoir-faire au sein de la DGFiP ;
4. L'accès aux données au travers d'APIM reste soumis à un processus en 2 étapes (données fictives puis données réelles) ce qui constitue un frein au recours aux API pour les partenaires et constitue une surcharge d'instruction ;
5. La stratégie d'homologation SSI retenue n'est pas aboutie à ce jour. Si une homologation s'appuyant sur une analyse de risques apparait bien prévue dans le projet Datalake, un retard dans cette démarche est susceptible de compromettre la bonne prise en compte de la sécurité dans le projet et de modifier en profondeur l'architecture retenue du fait de la sensibilité des données. Par ailleurs, la stratégie d'homologation montre que le métier, et en particulier la maîtrise d'ouvrage Cap usager, n'est pas invitée à la commission. L'autorité d'homologation et tous les membres de la commission font partie du service des systèmes d'information. La maîtrise des risques est abordée sous un prisme technique et pourrait par conséquent être en décalage avec les besoins opérationnels des métiers ;
6. Certains coûts du projet n'apparaissent pas dans le chiffrage du projet, nuançant la rentabilité affichée du projet, notamment la prise en compte détaillée de l'alternative à Cloudera, coûts de formation, de sécurité….

Les besoins étant avérés, les premiers résultats prometteurs, j'émets **un avis conforme positif sur le projet « Datalake et APIM » et donne mon accord au déblocage des fonds prévus au contrat FTAP, soit 8 219 500 € (CP)**.

Je vous encourage néanmoins, à prendre en compte les recommandations suivantes :

- Clarifier et intégrer à la trajectoire du projet l'impact technique, calendaire et budgétaire de la fin programmée de la maintenance d'Hortonworks.
- Renforcer et mettre en œuvre les mesures de promotion et de conduite du changement du programme permettant de sécuriser l'adoption à la fois du Datalake et des APIs auprès des métiers et des partenaires externes.
- Si la DGFIP a déjà mis en œuvre des actions limitant le risque de dépendance à des compétences externes (un plan de formation MOE, mise en pratique d'un POC et d'un « bac à sable », sous monitorat de compétences externes, recrutements…), ce risque demeure néanmoins et le plan de recrutement du projet apparaît très ambitieux s'agissant de ressources rares.
- Opérationnaliser la mutualisation de ressources et le partage des compétences entre projets mobilisant des techniques scientifiques élaborées de type « Data Science » au sein de l'Etat apparaît indispensable. Outre CVFR (DGFiP) et Intelligence Emploi (Pôle Emploi), une attention particulière devrait être accordée au développement d'un partenariat avec le projet 3D Douane, conduit au sein du même ministère.
- Permettre un accès ouvert et immédiat aux données fictives au travers d'APIM, sans avoir à répondre à un processus d'habilitation. Il convient également de viser une instruction des demandes d'accès aux API en moins de 10 jours.
- Compléter sans délai la stratégie d'homologation SSI pour pouvoir ensuite démarrer en priorité l'analyse de risques attendue. Le métier devra également être impliqué dans les travaux d'homologation.

Conformément au décret n° 2019-1088 du 25 octobre 2019 relatif au système d'information et de communication de l'État, la transmission du présent avis met fin à la procédure de saisine.

Nadi BOU HANNA

Directeur interministériel du Numérique

Copies :

Monsieur le Premier ministre
A l'attention de :

- Monsieur le directeur de cabinet

Madame la ministre de la transformation et de la fonction publiques
A l'attention de :

- Monsieur le directeur de cabinet
- Monsieur le directeur interministériel de la transformation publique